DOCUMENT RESUME

EL 155 204 TM 007 129

AUTHOR Tallmadge, G. Kasten; Horst, Donald P.
TITLE The Use of Different Achievement Tests in the ESEA Title I Evaluation System.
PUB DATE Mar 78
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Asscciation (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Analysis; Compensatory Education Programs; Content Analysis; Criterion Referenced Tests; *Curriculum; Curriculum Evaluation; Curriculum Research; Elementary Secondary Education; Equated Scores; Measurement Goals; National Norms; Norm Referenced Tests; *Program Evaluation; Research Design; Standardized Tests; Testing Problems; *Test Selection; *Test Validity; Validity
IDENTIFIERS *Content Validity; *Elementary Secondary Education Act Title I

ABSTRACT

In this discussion of the use of achievement tests in evaluating Title I programs, matching content between test and curriculum is the main theme. The point is made that unless a test measures what is taught, it cannot be sensitive to whatever gains the instruction produces. Thus, if different instructional treatments have different objectives, it is unlikely that a single test will provide equally valid assessments of their effects. Where different tests are needed to assess the effects of different curricula, national norms can provide a common metric for comparing gains measured with different tests. A gain of one-third of a national standard deviation on Test X, for example, would be considered equal in value to a gain of the same size on Test Y assuming that the two tests were equally good matches to the corresponding curricula. This approach will work quite well, presumably, when "typical" standardized reading or mathematics achievement tests are used. It breaks down if both test and curriculum are very narrowly focused. Under these circumstances gains will appear too large.
(Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED155204

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Donald P. Horst

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

THE USE OF DIFFERENT ACHIEVEMENT TESTS IN THE ESEA TITLE I EVALUATION SYSTEM

G. Kasten Tallmadge and Donald P. Horst
RMC Research Corporation

March 1978

Paper presented at the
Annual Meeting of the American Educational Research Association
Toronto, Quebec, Canada, March 27-31, 1978

TM007 129

One of the fundamental rules of good experimental design is that, if one wishes to compare different treatments, one must use a single yardstick. The Title I evaluation and reporting system, on the other hand, permits LEAs to use almost any test they wish. Some evaluators have suggested that the use of widely discrepant tests completely invalidates the system. Others have viewed this source of variance as a necessary evil, dictated by the realities of State and local autonomy. The title of our paper, in the context of this symposium,[1] might suggest a discussion of how we justify the imprecision introduced by aggregating data from different tests. In fact, we wish to defend a very different position on the instrumentation issue. While not denying the problems inherent in the use of many different tests, we are convinced that, compared to the alternative of a single, system-wide test battery, allowing LEAs to select their own tests substantially <u>increases</u> the precision of the system.

Background

The Title I evaluation and reporting system was designed to address the question, "Are Title I students benefitting from the service they are receiving?" This question was posed at the Federal level, but the relevant legislation makes it clear that the answer is to come from State and local level evaluations. Section 151 of P.L. 93-380 stipulates that:

> (d) The Commissioner shall provide to State educational agencies models for evaluations of all programs conducted under this title . . . , which shall include uniform procedures and criteria to be utilized by local educational agencies, as well as by the State agency in the evaluation of such programs.

It goes on to specify:

> (f) The models developed by the Commissioner shall specify objective criteria which shall be utilized in the evaluation of all programs and shall outline techniques . . . and methodology . . . for producing data which are comparable on a statewide and nationwide basis.

---

[1] <u>Sources of Imprecision in the USOE ESEA Title I Evaluation and Reporting System</u>.

Thus, Title I evaluations are to be conducted at State and local levels, but they are to yield results that are comparable and that can be interpreted on a nationwide basis.

In attempting to meet this legislative mandate we discussed the problem with 54 State Title I Coordinators and/or their designated evaluators. One of the questions we asked all of them was

> Would it be feasible to impose standardized evaluation practices on LEAs for the assessment of achievement benefits, including the specification of acceptable standardized test instruments and testing times?

The response to that question was somewhat less than overwhelmingly enthusiastic. Nevertheless, most of the States acknowledged that some degree of standardization would be acceptable. A few felt it would be highly desirable. On the other hand, even where the response to standardization was positive at a general or theoretical level, there was a good deal of hesitancy with respect to specific details. Particularly where States had taken a strong leadership role in standardizing their own practices, we encountered substantial resistance to change--not to standardization, but to change. In other words, such States felt, not unexpectedly, that the practices <u>they</u> had developed should be adopted nationwide.

This sentiment was particularly strong in those States that had adopted or were considering adopting criterion-referenced testing. Interviewees made it clear to us that they would regard any mandate to employ standardized achievement tests as a return to the dark ages. They strongly encouraged us to shun such retrogressive thinking.

Other States, of course, were equally committed to norm-referenced evaluations using particular standardized tests.

The Case Against a Single Test Battery

By the time we had completed our telephone interviews with the Title I coordinators, it had become clear to us that an ideal evaluation system would have to meet the legislatively mandated requirement of making individual State and local evaluations comparable while, at the same time, allowing local options with respect at least to the matter of test selection. These two design objectives, unfortunately, appeared to be incompatible.

Exact comparability of results across different evaluations, it seemed to us, would require that all of them employ the same instruments, or that the selection of instruments at least be limited to a few that could be carefully equated. Designing the system in that way, however, would have alienated a large segment of the population from whom we would be seeking cooperation. It appeared that we faced a true dilemma.

Gradually, as we continued to work on the system, our position changed. We now believe that a strong technical case can be made for test selection at the local level and that designing this feature into the system actually increases the validity of local, State, and national Title I evaluations. Our belief is based on both empirical data and theoretical arguments, and we would like to begin this morning by reviewing some of the empirical evidence that has led to our current position.

Many of us have often been involved, one way or another, with local-level evaluations that failed to show the dramatic achievement benefits claimed for the programs being evaluated. It has usually been our experience in these situations that some program person would dismiss the negative findings with a statement something like, "Well, that was to be expected. The test doesn't measure what the program is accomplishing!"

We should point out that we are quite skeptical evaluators and do not accept such explanations lightly. In at least a few instances, however, our own observations have led us to believe that students were being

substantially benefitted by special instructional programs and we have been shocked when the benefits failed to show up on the tests. As a result of these experiences, we began to entertain the hypothesis that not all achievement tests are sensitive to all achievement gains.

We now know that that position has been attracting an increasing number of advocates in recent years. In commenting on the negative findings of a very carefully executed evaluation of several presumably exemplary compensatory education projects, Kaskowitz and Norwood (1977), for example, had this to say:

> The demonstrated correspondence of test content with program or policy objectives is of the utmost importance. If such correspondence cannot be shown, we feel that no basis for the use of standardized tests exists.

And, indeed, after examining the content of the Metropolitan Achievement Test they found that it was a poor match with the content of any of the projects they had attempted to evaluate.

A paper presented by Ralph Tyler at the 1976 U.S. Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Program Evaluation made a similar point. He emphasized that there are a number of different curricular routes that may be followed to attain the same reading or math objective. He used the example of reading comprehension, pointing out that the acquisition of that skill generally takes place over a number of years. Thus, he argued that the evaluator must either allow enough time for the entire instructional sequence to take place or he must tailor his testing to that segment of it that occurs between pre- and posttesting.

This argument seems particularly relevant when considering the distinction between policy objectives and program objectives. While it may be a policy objective of Title I to enhance the ability of children to read in the global sense that that ability is operationally defined by

standardized reading tests, individual instructional programs are necessarily focused on some subset of component skills. It may then be the case that the component skills taught in a particular reading program will not even be sampled by the standardized test selected to reflect the overall policy objective.

Ralph Hoepfner of the System Development Corporation has taken the position (1976) that it is the global kind of policy objectives that are most relevant to an evaluation such as that which SDC is undertaking of the sustaining effects of compensatory education. That evaluation, however, is a longitudinal study spanning several years. One suspects that the information he himself compiled in selecting a test for that study might have led him to adopt a different stance for shorter term evaluations.

He categorized test item frequencies according to the instructional objective they measured and prepared tables comparing the content coverage of the eight standardized achievement tests included in the Anchor Test Study. Let us highlight some of his findings. The reading portions of the third-grade level of both the California Achievement Test and the Cooperative Primary devote approximately 30% of their items to word attack skills. Of the remaining six tests, not a single item was devoted to word attack skills. At sixth grade, the California and Metropolitan Achievement Tests devoted about 50% of their items to the recognition of word meanings. According to Hoepfner's analysis, no items covering this objective were included in the Sequential Tests of Educational Progress or the Stanford Achievement Test.

The percentage of items devoted to reading comprehension ranged across tests from 14% to 53% in first grade, from 17% to 50% in third grade, and from 25% to 64% in fifth. Reading interpretation varied from 0% to 38% in both fourth and fifth grades.

The situation appeared somewhat better in mathematics but he found some notable differences there too. Whole-number computation, for example,

accounts for only 2% of the items at the fourth-grade level of the Iowa Test of Basic Skills, while 62% of the items on the California Achievement Test are devoted to this objective. It is not hard to imagine how this difference would affect the results of an evaluation of a remedial arithmetic program! Yet we fear that tests are often selected without considering such matters.

A small but very interesting and relevant study was recently carried out by Armbruster, Stevens, and Rosenshine (1977) at the new Center for the Study of Reading at the University of Illinois. They categorized the content of three reading curricula designed for the second half of third grade (the Economy Company's Mysterious Wisteria, the Ginn 360 Reading Series All Sorts of Things, and Houghton Mifflin's Fiesta) and two widely used standardized tests of reading comprehension (California Achievement Test and Metropolitan Achievement Test). They used a 16-category classification system borrowed from SRA's SOBAR Reading Catalog of Objectives 3-9 (1975).

The percents of all curriculum exercises falling into each category were computed. Then the categories were rank-ordered within curricula according to these percents. Rank-order intercorrelations were calculated among the three curricula based on all 16 categories of objectives with the following results. The Economy and Houghton Mifflin curricula intercorrelated .40, Economy and Ginn .31, and Ginn and Houghton Mifflin .33. These low correlations were felt to reflect important differences in emphasis among the curricula.

All 16 reading comprehension categories were covered in one or more of the curricula. However, only 6 were found to be included in the coverage of the MAT and 7 in the coverage of the CAT. The authors figured that 64% of the Economy, 65% of the Ginn, and 79% of the Houghton Mifflin reading comprehension exercises were not covered at all by MAT or CAT test items.

Using just the 6 categories that all three curricula and both tests had in common, intercorrelations among the curricula were recalculated.

They ranged from -.34 (Houghton Mifflin and Ginn) to +.43 (Houghton Mifflin and Economy). Correlations between the tests and the curricula based on the same six categories ranged from .10 (MAT with Ginn) to .43 (CAT with Houghton Mifflin).

It seems highly probable that, where the content of a test shows a low correlation with the content of a curriculum, the test will be insensitive to whatever gains the curriculum might produce. The problem is aggravated by the fact that students gain only a few raw score points on a total standardized test during a normal school year. If only a few sections of a test are relevant to the curriculum, even dramatic gains on these sections may have little impact on total scores.

Unfortunately, empirical data on the effects of test relevance are hard to find. One interesting study is that by Bianchini (1976). He reported that when statewide testing of California first graders was done with the Stanford Achievement Test, mean performance was found to be at the 38th percentile of the national norms. The following year, when a switch was made to the Cooperative Primary, mean performance was almost exactly at the 50th percentile. Although data from the Anchor Test Study suggest that the norms of that edition of the SAT were probably "harder" than those of the Cooperative Primary, the principal cause of the shift was judged to be the improved match between test content and curriculum content. Bianchini found that only 19% of the vocabulary in the SAT was contained in the instructional materials provided by the State. He found a 55% correspondence with the vocabulary of the Cooperative Primary.

Implications for Comparisons Across Projects

Other studies are currently underway that should shed further light on this issue. We suspect, however, that most of us already have a pretty good idea what these studies will find. We feel confident enough so that we have been strongly recommending that evaluators carefully examine the content of all tests that might be candidates for use in particular situations and that they consider the match with curriculum content the overriding determiner of the ultimate selection.

It seems to us that the only valid way to assess the effects of an instructional treatment is to use a test that measures what was taught. Not item by item of course--but the test items should be samples from the same domains as the teaching/learning exercises. It further seems to us that any evaluation that does not measure what was taught can only yield results that are invalid. Finally, because different Title I reading and math projects are teaching different subsets of the components of those overall abilities, it seems unlikely that any single reading or math test exists or could be developed that would be appropriately sensitive to the effects of the entire range of treatments.

The implications of this situation for a Title I evaluation system that is supposed to yield comparable results are not entirely clear. If comparability means measuring the same thing in all circumstances, regardless of whether or not it is appropriate, then, to achieve comparability, we would be forced to use a single test or a small number of similar, carefully equated instruments. Under these circumstances, however, we would generate large amounts of data that did not accurately reflect the outcomes of the treatments being evaluated. Suppose, for example, that a truly outstanding program for teaching whole-number computations to fourth graders ended up being evaluated by the ITBS. Since only 2% of the test items tap this skill, there is no chance that the test results would reflect the true impact of the treatment. They could easily show that the treatment had no impact whatsoever, or perhaps even that it had a negative effect on the participating students.

It seems that if we define comparability in terms of content we will end up with a situation in which highly effective projects may appear less beneficial than projects that are only moderately effective. Furthermore, if this occurs, our overall assessment of the Title I program will be systematically distorted (and we suspect seriously so) in a negative direction. The situation would be something like evaluating all world-class athletes on a test of bowling. Not only might we find that Laura Baugh scored higher than Nadia Comaneci, we would also find that, on the average, world-class athletes are really pretty mediocre performers.

What Do We Really Want to Measure?

We believe that there is an important dilemma reflected in this analogy and that it is not subject to resolution via better experimental designs or more elaborate statistical treatments. The dilemma concerns the nature of the gain that is measured and its relationship to students' overall cognitive growth. Consider, for example, a highly focused instructional treatment evaluated with a totally relevant test. Unless the control, comparison, or norm group had received similar instruction, the evaluation would almost certainly show that the treatment had a very large impact. On the other hand, that evaluation outcome seems unfair. It could certainly be argued that a smaller gain measured over a broader content area was at least equally valuable to the students.

The ultimate aim of Title I reading projects is usually to improve reading in a global sense. A particular project, however, may focus only on a few component skills and, as suggested earlier, may cover only a selected subset of these skills in any given year. Ideally, improvement in one set of component skills would be followed by improvements in other skills. Thus, what is learned one year would carry forward to the next and lead ultimately to the desired improvement in reading in the more global sense. Unfortunately, there is as yet little evidence to support such an assumption.

A closely related issue is whether the student learns <u>additional</u> skills because of the project or simply skills that are different from what would have been learned without the project. Obviously we want the students to learn <u>additional</u> content but, before we assume that this has happened, we need to consider several other possibilities.

The first possibility is that, rather than providing <u>more effective</u> instruction, the project simply provides <u>more</u> instruction. If the project simply provides additional "time on task," the time has to come from somewhere, and it is likely that that the student will learn less about other specific items, component skills, or subject areas. A second possibility

is that the total amount of time spent on each subject before the completion of schooling is unchanged but the sequence of the curriculum is altered for Title I students so that, in any given time period, they do not cover the same content as their non-Title I peers. In this case, the project may look either good or bad depending on the particular time period under consideration, the test that is used, and the comparison group that is chosen, while, in fact, there may be no impact at all in the long run.

To summarize, although we may hope or even believe that our project students are maintaining their normal growth in other skill areas while receiving special help in reading, it may be the case that measured growth in selected reading skills is actually achieved at the expense of other valued skills or knowledge. If the project students learn anything <u>different</u> from that which is learned by the norm or comparison group that provides their no-treatment expectation, the apparent effectiveness of the project will hinge on the extent to which the content of the criterion test matches the content of one or the other of the curricula. Thus, we can control the apparent impact of a project to a great extent by the test that we chose. The question is, exactly how much value should be placed on improvements in a restricted range of skills? A test that samples all the skills of the comparison group may fail to detect a real impact of a Title I project. A test tailored to the project may give an unreasonably favorable impression.

A Practical Solution to the Problem

The above line of argument has convinced us that no neat and tidy, completely satisfactory answers will be found to the problems of which tests to use and what interpretations to place on the results. In practice however, the situation may not be as bleak as we have just painted it. As so often happens when logical arguments lead us to reject all solutions, a limited amount of compromise with our theoretical principles allows us to deal reasonably well with reality.

To illustrate such an approach, let us return to the example of evaluating athletes. If a test of bowling is an inappropriate mechanism for

selecting the athlete of the year--or the most improved athlete, to make the analogy a little more relevant--how should a reasonable selection be made? Certainly, there is no entirely satisfactory solution to the problem. On the other hand, if one is willing to make the assumption that all sports are equal, a fairly reasonable approach becomes possible. Given that equality, a ski jumper who is 3 standard deviations above the mean of competative ski jumpers can be considered a better athlete than a high jumper who is only 2.9 standard deviations above the mean of his competative peers. Similarly, a weight lifter who has increased the amount he can clean and jerk by a quarter of a standard deviation is less improved than a marathon runner who has reduced her running time by a third of a standard deviation.

The process we have just described is somewhat analogous to the manner in which the Title I evaluation and reporting system currently attempts to make evaluations of diverse treatments yield comparable results. The system measures gains in terms of national standard deviation units and assumes that a gain of, say, a third of a standard deviation on any one test is as good as a gain of a third of a standard deviation on any other test. This assumption of equivalence was never meant to imply that if a Title I group showed a third of a standard deviation gain on one achievement test, they would show precisely the same size gain on another. That would be like assuming that a player's improvement in tennis would automatically be accompanied by an exactly equal-sized improvement in every other athletic skill. All the system implies is that a project that produces a gain of given size on one test is as effective in meeting the objectives of the Title I program as another project that produces a gain of the same magnitude on some other test.

Admittedly there are flaws in this assumption, flaws that will make the data from the Title I evaluation and reporting system more ambiguous than we would like. In the first place, no attempt is made to measure the effect of the Title I projects on skills or subject areas outside of the immediate scope of the projects themselves. Then too, there is pressure to show good results from Title I projects, and careful matching of

content and tests may make projects look better than they really are. Model A2, which permits criterion referenced and locally developed tests, is particularly susceptible to this problem. On the other hand, Model A1, which involves the use of standardized tests, probably tends to be conservative, since standardized test items will usually not match the content of most projects at the level of specific items. Even in control group designs with truly random assignment to groups, the choice of tests will favor one of the groups relative to the other.

It is obvious that if some tests match content too well and others don't match well enough then, in theory, there should be tests that are in between these extremes and match exactly right in terms of the conclusions we wish to draw. At this time, we know of no way to identify or develop such perfect tests. Our feeling is, though, that for most reasonably conventional projects, a carefully selected standardized test is a fairly good approximation to the ideal test. Where results seem far too good or far too poor to be believable, the use of a second standardized test provides an easy check on student progress. In any case, we are firmly convinced that the system as it is currently designed, will yield a much fairer assessment of what Title I is accomplishing than could be obtained using a single test to assess all treatments.

REFERENCES

Armbruster, B. B., Stevens, R. J., & Rosenshine, B. Analyzing content coverage and emphasis: A study of three curricula and two tests. Urbana, Illinois: University of Illinois, 1977. (Technical Report No. 26)

Bianchini, J. C. Achievement tests and differentiated norms. Paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, Virginia, 1976.

Hoepfner, R. Achievement test selection for program evaluation. Paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, Virginia, 1976.

Kaskowitz, D. H., & Norwood, C. R. A study of the norm-referenced procedure for evaluating project effectiveness as applied in the evaluation of Project Information Packages. Menlo Park, California: SRI International, 1977. (Research Memorandum, Contract OEC-0-74-9256)

Tyler, R. W. Discussion of Hoepfner's Paper. Paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, Virginia, 1976.